CRET（教育テスト研究センター）研究会
2011.11.24

妥当性と信頼性の数理

テストの開発・評価への応用の視点から

南風原朝和
（東京大学大学院教育学研究科）

1

概　要

妥当性と信頼性を統一的なモデルで表現し，それに基づき，以下の事項について解説する。

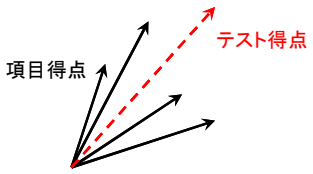
- 妥当性と信頼性の相互関係
- 系統誤差とグローバルおよびローカルな偶然誤差
- 妥当性と信頼性を高めるために必要なこと
- 信頼性の諸指標が意味するもの
- テストの開発・評価で用いられる手続きの見直し
- 妥当性検討の実際例
- テストの開発・評価で留意すべき点

2

想定するテスト

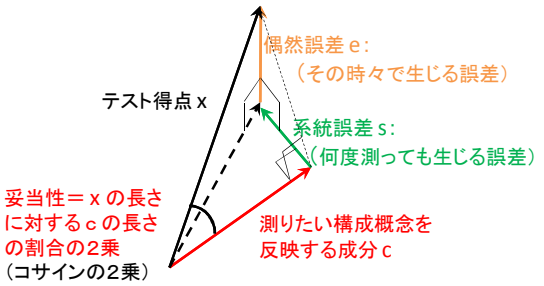
主として認知的なテストを想定（一部，質問紙尺度も）

複数の項目から構成され，項目得点の和（または平均）をテスト得点とするもの



3

テスト得点の成り立ちと妥当性



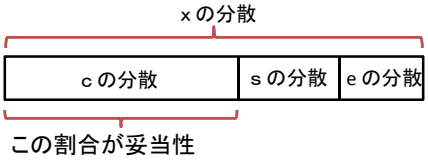
4

テスト得点の成り立ちと妥当性

テスト得点（x）　＝

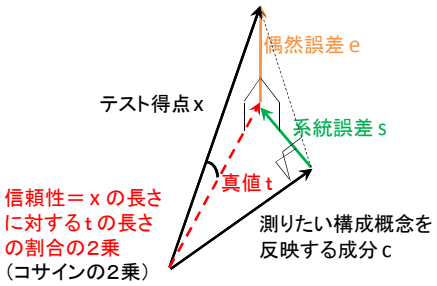
測りたい構成概念を反映する成分（c）

＋　系統誤差（s）＋　偶然誤差（e）



5

テスト得点の成り立ちと信頼性

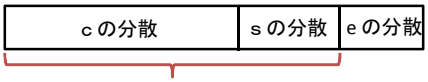


6

テスト得点の成り立ちと信頼性

テスト得点（x） ＝

真値（t）［測りたい構成概念を反映する成分（c）
＋ 系統誤差（s）
＋ 偶然誤差（e）



この割合が信頼性

7

妥当性と信頼性の相互関係

・信頼性が低いと，妥当性は高くなりえない

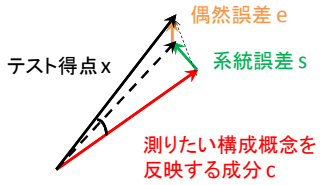
・妥当性が高ければ，信頼性は高くなる

・信頼性が高くても，妥当性は低い可能性がある

⇒ 高い信頼性は高い妥当性のための必要条件であるが，十分条件ではない

8

テスト作成における目標



s と e をcに比して小さくすることによって，x の向き（テスト得点のベクトル）を c の向き（構成概念のベクトル）にできるだけ一致させること

9

テスト作成における目標

テスト得点の妥当性を高くすること

＝ 系統誤差 s の分散を小さくすること
＋偶然誤差 e の分散を小さくすること
＋構成概念を反映する成分 c の分散を大きくすること

このうち第2項は信頼性の問題

10

系統誤差と偶然誤差

テスト得点に含まれる系統誤差と偶然誤差は，それぞれ項目得点に含まれる系統誤差および偶然誤差の和（または平均）

・項目得点に含まれる偶然誤差
⇒ “一般には”項目数が多ければ相殺される
（高い信頼性のためには多くの項目が必要）

・項目得点に含まれる系統誤差
⇒ 項目数が多くても相殺されるとは限らない

11

偶然誤差の要因

ローカルな要因とグローバルな要因の区別

ローカルな要因
＝ 項目ごとに独立に作用する要因

例：回答の際の一瞬の迷い
項目の読み間違い
回答のマークのつけ間違い

・項目数を多くすると誤差が相殺される

12

偶然誤差の要因

グローバルな要因
＝ 項目全体に作用する要因

例：テスト当日の体調の悪さや気分

・偶然変動だが，項目全体に影響
・項目数を多くしても誤差が相殺されない
・項目間の相関においては真値の一部のように機能し，項目間の相関に寄与するため，α係数の値は不当に高くなる
・誤差が相殺されるためには複数機会での測定が必要

13

グローバルな要因による偶然誤差と項目間相関

項目jの得点xj＝項目jの“本来の”真値tj
＋グローバルな要因による偶然誤差gj
＋ローカルな要因による偶然誤差lj
（tj，gj，lj間は無相関と仮定）

Cov(xj, xk) = Cov(tj, tk) + Cov(gj, gk) + Cov(lj,lk)
（＞0） （＝0）

gjとgkが相関する分，項目間相関が高まる
⇒ 因子分析やSEMでも誤差として分離されず，
共通因子に含まれることになる （⇒ 実力のうち？）

14

偶然誤差を減らすには

ローカルな誤差
・読み間違いのない文章
・マークのつけ間違いのない回答フォーマット
・実施環境の調整，回答への動機づけ
・項目数を増やす

グローバルな誤差
・実施環境の調整，回答への動機づけ
・（可能なら）測定機会を増やす

15

信頼性の推定値に反映される誤差

α係数および折半法信頼性：
ローカルな偶然誤差

再検査信頼性：
ローカルな偶然誤差
グローバルな偶然誤差

※ 項目の等質性（内的一貫性）は，α係数だけでなく，再検査信頼性にも反映される

16

再検査信頼性のしくみ

項目jの得点xj＝項目jの“本来の”真値tj
＋グローバルな要因による偶然誤差gj
＋ローカルな要因による偶然誤差lj （tj，gj，lj間は無相関と仮定）
テスト得点の分散（各回同じ，再検査信頼性の分母）
＝ Σtjの分散 ＋ Σgjの分散 ＋ Σljの分散
1回目と2回目のテスト得点間の共分散（再検査信頼性の分子）
＝ 1回目のxjと2回目のxkの総当たりの共分散の総和
＝ 1回目のtjと2回目のtk（=1回目と同じ）の共分散の総和
（gj, lj は回の間で相関しないため）
∴ 再検査信頼性はgjの分散共分散を反映して低下し，
項目真値tj間の相関（項目の等質性）を反映して高まる

17

系統誤差の要因

偶然誤差以外で，構成概念に無関係な分散（construct irrelevant variance）を生じさせるものすべて

項目の内容的・方法的偏り（construct underrepresentation）
構成概念の内容の一部が反映されていなかったり，
バランスが悪かったり，測定の方法が偏っていたりすること

（例）数学のテストの項目が，代数領域に偏り，幾何領域が少ない ⇒ 構成概念ベクトルとのずれ増大
（数学能力の「タイプの違い」による無関係な分散が増大）
⇒ 妥当性低下 （等質性が高まり信頼性は向上）

いわゆる内容的妥当性の問題も「無関係な分散」で説明できる

18

系統誤差の要因

質問紙尺度などでは，

反応バイアス
社会的望ましさ，黙従傾向など

別の構成概念を反映：（例）自己評価の甘さ
測定される構成概念のレベルが同じでも，自己評価の甘い人は，ポジティブな特性については高く，ネガティブな特性については低い得点に

19

系統誤差を減らすには

系統誤差（＝構成概念に無関係な分散）を減らすには，さまざまな個人差のうち，構成概念に照らして何がrelevantで何がirrelevantかを判断することが必要

そのためには，測定したい構成概念を明確に定義し，それが何を含み何を含まないのか，含むものはどのようなバランスで含むのかを示すことが必要

↓

「テスト仕様書」（構成概念の定義，要素や下位領域，対象者の範囲，項目形式等）の作成
＝構成概念のベクトルを定位する作業

テスト仕様書は，妥当性を高めるためにも，妥当性を検討する際にも，またユーザがテストを選択するうえでも有用

20

系統誤差を減らすには

例1：項目の内容的偏り
⇒ テスト仕様書にもとづいて項目を作成・選択
不足内容の追加，重複内容の削除

例2：別の概念を反映
⇒ 予備テストにより，その概念のテストとの相関が特に高い項目を削除または改訂

21

テストの開発・評価で用いられる手続きの見直し

ルーティン化された以下の手続きについて，ここまで述べた妥当性の向上（系統誤差・偶然誤差の低減）の観点から，どのような効果があるか検討

1．項目分析
項目-全体相関
天井効果・床効果，分布の歪み
2．探索的因子分析
3．他の変数との相関による妥当性検討

22

項目分析（項目-全体相関）

1次元的なテストを作成する際に，項目全体の合計点と各項目の得点との相関を調べ，相関の高いものを「識別力」の高い項目として選択する手続き

ここでの「識別力」は合計点の高低を識別する力 ⇒ 合計点のベクトルが構成概念のベクトルに近似している場合には，妥当性を高める手続きとなる

合計点がほぼ妥当なテスト得点とみなせるときに，その方向からずれている項目をチェックする手続き ⇒ その条件が満たされないときは，より妥当な項目が残る保証はない

1因子解の負荷やIRTの識別力を用いる場合も同様で，因子やθ尺度と構成概念のベクトルの向きが少なくとも近似的に一致していることが前提（探索的因子分析の項で詳述）

23

項目分析（天井効果・床効果）

たとえば5段階評定の項目において，5や4（あるいは1や2）に回答が集中し，個人差の識別が十分にできない項目をチェックし，削除ないしは改訂する手続き

偏りがあっても，構成概念を反映する良い項目である可能性があるので，機械的な判断は避けるべき

一部で用いられている「平均±SDが評点の端（5段階なら5と1）を超えたら天井（床）効果とする」という手続きは，平均が同じなら個人差（SD）の大きい項目のほうを除外することになり，妥当とはいえない

24

項目分析（分布の歪み）

項目得点の分布の偏りは，分布の歪み（正規分布からの逸脱）として問題にされるケースも多い

変数の正規性を仮定した因子分析（最尤法）を用いる際の仮定のチェックとしては意味があるが，因子分析に含めることの可否と構成概念の測度としての有用性の有無は別問題

項目得点が単独でではなく合計されてテスト得点に組み入れられることからも，項目得点の（妥当性以外の）統計的特徴に過敏になる必要はない

25

探索的因子分析

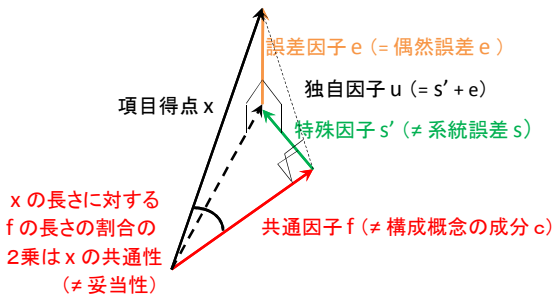
「テスト作成といえば（探索的）因子分析」というくらい，結びつきが強い

「同一の構成概念を測るための項目は相互に相関が高く，異なる構成概念を測るための項目は相互に相関が低い」というのは妥当性のための要請であり，因子分析はその要請が満たされているかどうかを調べるには有用

ただし，因子が構成概念と同一視されてしまうと，妥当性の観点から問題が生じる

26

因子分析のモデル



27

共通因子と構成概念

一般に，同一の系統誤差要因やグローバルな偶然誤差要因が多数の項目に影響を与える可能性がある
⇒ 大多数の項目が構成概念の向きから大きく外れる可能性がある（＝現実的状況）

一方，項目の特殊因子は，系統誤差やグローバルな偶然誤差とは異なり，他の項目の特殊因子と無相関であると仮定される（いわば“ローカルな系統誤差”）
⇒ 項目のクラスターが共通因子の方向から大きく外れることはなく，共通因子（2因子以上の場合は共通因子平面）はつねにクラスターの中心に

したがって，構成概念のベクトルと共通因子のベクトルは一般には一致せず，これは項目数を増やしても解消されない

28

探索的因子分析と妥当性

妥当性の観点からは，因子負荷（やIRTの識別力）をもとに項目選択をするのは，因子（や潜在特性）と構成概念のベクトルの向きが少なくとも近似的に一致していることが前提

バリマックスやプロマックスの単純構造の原理だけで探索的に回転して得られる因子，しかも回転法によって変化する因子が，この前提を満たすか
⇒ その方法論的根拠はない

29

探索的因子分析と確認的因子分析

構成概念の定義を含むテスト仕様書作成から，妥当性の高いテスト作成を目指す流れにおいては，探索的因子分析より確認的（検証的）因子分析がより適している

構成概念の定義やテスト仕様書の作成が困難な，文字通り探索的な段階であれば，項目のグルーピングや，構成概念についての検討を始めるための探索的因子分析は有用

しかし，構成概念の定義が困難であった段階で用意した項目セットをもとにした探索的な分析で，一挙にテスト化まで進むのは無理がある（構成概念の定義と，そのもとでの項目の作成・選択のステップが必要）

30

既存のテストの探索的因子分析

既存のテストを使用する際に探索的因子分析を適用し，その結果がテスト作成者が示した因子構造と異なる場合に，テストを組み換えて分析するケースがある

少なくとも以下の点の検討が必要

- 先行研究で得られた因子構造の確認のために探索的アプローチを用いるのは適切か
- 「標本変動の範囲内のずれ」という説明は棄却できるか ⇒ 負荷の標準誤差，多群解析
- 因子構造が異なることの合理的説明は可能か

31

因子負荷の標準誤差，信頼区間，被覆表示（SASの例）

Rotation Method: Promax (power = 3)
Rotated Factor Pattern (Standardized Regression Coefficients)
With 90% confidence limits; Cover |*| = 0.45?
Estimate/StdErr/LowerCL/UpperCL/Coverage Display

	Factor1	Factor2	Factor3
test8	0.90139	-0.09763	-0.02002
標準誤差 ➡	0.03568	0.03664	0.03506
信頼区間 ➡	0.82317	-0.15746	-0.07756
	0.94603	-0.03708	0.03766
被覆表示 ➡	0*[]	*[]0	*[0]

32

確認的因子分析モデルの選択

1つの構成概念に1つの因子を対応させる場合，単純な1因子モデル（独自因子は互いに無相関）をあてはめると，項目数がある程度多くなると，適合が悪くなる

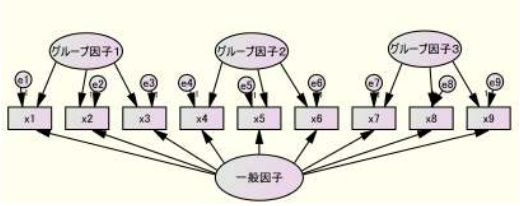
構成概念の定義によって，

- 構成概念に対応する「一般因子」と，その中の各下位領域に対応する「グループ因子」からなる階層因子分析モデル
- 下位領域ごとの「1次因子」をまとめる「2次因子」を想定する2次因子分析モデル

などのモデルを柔軟に選択

33

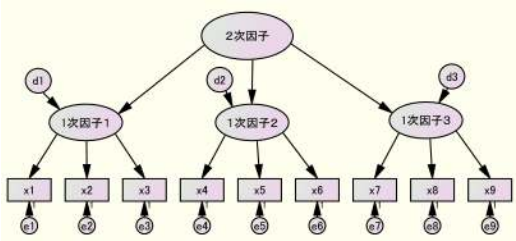
階層因子分析モデル



・グループ因子は一般因子と無相関
⇒ 残差間の相関をまとめる役割

34

2次因子分析モデル



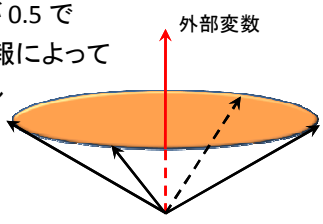
・1次因子は2次因子を共通に反映
⇒ 特殊な部分は残差変数 d に

35

他の変数との相関による妥当性検討

相関が予想される2，3の変数との相関を求めることで，妥当性の確認とされることが多い

外部変数との相関が 0.5 であった場合，その情報によってテスト得点のベクトルの向きはどの程度，定まるか　⇒ 図



36

他の変数との相関による妥当性検討

外部変数の数が増えるにしたがって，相関情報によるテスト得点ベクトルの定位（＝テスト得点の解釈の確立）が進められるが，2，3の変数との相関のみでは不十分

たとえば，弁別の必要がある関連テストの得点の影響を除いたうえで外部変数との相関（部分相関）を求めるなど，「妥当性への脅威」に直接応える工夫によって，意味のある妥当性情報が得られる

継続的な妥当性検討が必要

37

妥当性検討の実際例

Brackett & Mayer (2003)：　情動知能に関する競合するテスト（MSCEIT，EQ-i，SREIT）の妥当性検討

・それぞれからパーソナリティテスト得点（Big Five）と言語能力得点（verbal SAT）の影響を除いたうえで，日常生活での複数の基準変数を予測

・SREITの開発者は学校での成績を予測できるとしていたが，偏相関は負となり，他の変数とは有意な偏相関はなかった

・MSCEITと社会的逸脱，EQ-iと飲酒との相関は，偏相関でも残った

・“Most personality psychologists would agree that for a new construct to be welcomed into the field, it must explain variance that is not accounted for by well-established constructs.” (p.1156)

38

テストの開発・評価で留意すべき点

1．妥当性はテストの属性か

テスト得点の性質（ベクトルの向き）は，実施の仕方（時間制限など）を変えたり，本来の対象以外の集団に実施したりすると，変化する

また，構成概念の定義と異なるかたちでテスト得点を解釈する場合，その解釈の妥当性は，もとの構成概念とは異なるベクトルとの関係で評価される

その意味で，妥当性はテストそのものの性質というより，テストの使用に関する性質である

39

テストの開発・評価で留意すべき点

2．受験者のサンプリングの影響

妥当性も信頼性も，「真の部分の分散の割合」であり「真の部分と観測値との相関（の2乗）」である。

一般に，真の部分の分散は，選抜された集団，偏りのある集団では小さくなるが，誤差の部分の分散は変わらない。

したがって，選抜された集団，偏りのある集団でデータをとると，妥当性，信頼性とも見かけ上，低くなる。

40

テストの開発・評価で留意すべき点

3．受験者数の影響

妥当性も信頼性も，その推定値は統計量であるから，受験者数が少ないと標本変動が大きくなる

1回の実施で得られた結果を固定的にとらえないことが必要

4．スケーリングの影響

ベクトルモデルを初め，ここでは線形のモデル，線形の関係をベースにしたが，IRTの尺度値は通常のテスト得点とは非線形の関係にある。このことも相関に影響を与え，したがって妥当性・信頼性に影響を与える

41

まとめ：よりよいテストづくりのために

1．「妥当性」のとらえどころの無さからの脱却

⇒ 妥当性概念とテスト作成の目標の明確化

2．計画段階・作成段階からの「妥当化」

3．「テスト仕様書」が可能なら作成，
不可能なら作成に向けての探索的検討

4．偶然誤差・系統誤差の要因に関する検討とそれらを個々に抑制する工夫

42

まとめ：よりよいテストづくりのために

5. 項目分析は機械的にではなく，
   テスト作成の目標に合わせて
6. 因子分析は，研究段階に応じて，探索的
   アプローチと確認的アプローチの使い分け
7. 信頼性の推定は，査定したい誤差の種類に
   合わせて
8. 妥当性の検証は，多変数的なアプローチも

43

まとめ：よりよいテストづくりのために

テスト作成研究の論文評価

- 計画段階から作成段階を通して，
  「妥当化」のためのどのようなステップを
  踏んだか
- 「妥当性の証拠」として，どれだけ
  focused & risky な予測がデータで検証
  できたか

以上

44